Optimal Subtractive Dither for Near-Lossless Compression*

Matthew Klimesh†
klimesh@shannon.jpl.nasa.gov
Mail Stop 238-420, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109

Abstract

Subtractive dither is a technique which may be used to reduce the occurrence of compression artifacts from near-lossless compression. Standard subtractive dither incurs a cost, however, in the form of an increase in rate and distortion, and by giving the reconstructed signal an overall grainy appearance. It is possible to compromise between the costs and benefits of dithering by using dither signals which tend to take on smaller values than the standard dither signal. For reasonable cost and benefit metrics, the optimal dither signals are shown to be those which are uniform on intervals of the form $[-k/2, k/2]$, where $k \in [0, q]$ and q is the quantization step size. Additional results are given under more general assumptions.

1 Introduction

Most lossy data compression algorithms have some tendency to produce undesirable artificial features (artifacts) in the reconstructed data. We are primarily concerned with near-lossless compression of sampled signals, especially images. In near-lossless compression, the magnitude of the individual sample errors is usually strictly limited, but several correlated errors can produce artifacts which may interfere with scientific analysis of the signal, even if the distortion is normally almost imperceptible. These artifacts may include:

- A biased average value of some regions of the signal.
- "Contouring", or steplike signal value profiles, in slowly changing portions of the signal.
- Erasure of faint features which would be detectable in the original signal because they occupy a large area.

These artifacts are essentially caused by the fact that the quantization errors are not independent from each other or from the original signal values.

Subtractive dither [4, 5] is a technique for reducing or eliminating artifacts. In its simplest form, a pseudorandom dither value, uniformly distributed on $[-q/2, q/2]$

*See Section VII and Appendix C of [2] for a more complete version of this article.

†The work described was funded by the TMOD Technology Program and performed at the Jet Propulsion Laboratory, California Institute of Technology under contract with the National Aeronautics and Space Administration.

(where q is the step size of the uniform quantizer being used), is added to the value to be quantized. The same dither value is then subtracted from the quantized value during reconstruction. Thus in a predictive compression algorithm the reconstructed value is $\tilde{x} = Q_q(x - \hat{x} + D) + \hat{x} - D$, where x is the sample, $\hat{x}$ is the predicted value, D is the dither pseudorandom variable, and Q_q is a uniform quantization function with step size q. It is well-known that the resulting quantization noise $\tilde{x} - x$ is uniform on $[-q/2, q/2]$ and independent of x and $\hat{x}$.

When this dithering technique is used, the resulting root-mean-squared error (RMSE) distortion of the reconstructed signal will be exactly equal to $q/\sqrt{12}$, at least within the approximation that the dither signal may take on a continuum of values. This computation is based on the uniform error distribution. This RMSE value will usually represent an increase as compared to the RMSE for the same compression without dithering, since in the latter case the error distribution will generally be slightly peaked if the estimator producing $\hat{x}$ is performing well. The resulting rate (in bits/sample) will typically be higher also, since the values being quantized will have a larger variance. See [5] for some useful techniques for reducing these increases.

The reconstructed signal will have none of the artifacts mentioned above, since those artifacts occur due to correlations between the quantization noise and the signal. However, the entire signal may appear somewhat grainy due to the uniform noise on all samples.

It is possible to compromise between the dithering described above and no dither. This can be accomplished by using a dither signal which tends to take on values of smaller magnitude than would a signal uniformly distributed on $[-q/2, q/2]$. Such a dither signal will still increase the rate and distortion, but by a smaller amount, and it will remove some correlation between the reconstruction errors and the signal. However, many distributions are possible and it is not immediately obvious how to determine which are best.

In Section 2, we define metrics for measuring the relative costs and benefits of dither distributions. Based on these metrics, we then present results on the nature of optimal dither signals, including specific results which apply under reasonable assumptions.

2 Problem Formulation

We generally assume the original and reconstructed signal as well as the dither signal may take on a continuum of values; however, we later present some results for the discrete case.

We denote the probability measure of a general dither random variable D by P_D (so that $P_D(S) = \Pr(D \in S)$). Suppose $A(P_D)$ is a metric for the degree to which the reconstructed signal will contain artifacts, and $C(P_D)$ is a metric for the cost of the dither signal as manifested by the increase in rate and distortion. Then we would like to determine the dither signal distributions which achieve the optimal tradeoff between minimizing $A(P_D)$ and minimizing $C(P_D)$.

Two reasonable choices for $C(P_D)$ are the variance of D and the second moment

of D. (Note that these are the same if the dither distribution has mean 0.) Experimentally, the variance of D is a good indicator of the increase in distortion from dithering. It is logical that the variance of D also gives an indication of the increase in rate, since the rate is generally roughly equal to a constant plus the logarithm of the variance of the residual distribution. Figure 6(a) of [5] suggests that a similar relation will hold if the values of D are supplied to the entropy coder and decoder.

Reasonable choices of $A(P_D)$ are more complicated. When dithering is not used, the error $\tilde{x} - x$ in a reconstructed sample is dependent on the estimate $\hat{x}$ of the sample by $\tilde{x} - x = \hat{x} + Q_q(x - \hat{x}) - x$. When subtractive dither is used this error is random, and the dependence between the reconstruction error and the estimate takes the form of a possible "bias" in the reconstructed value. Specifically, $E[\tilde{x} - x] = E[Q_q(x - \hat{x} + D) - (x - \hat{x} + D)]$. Treating the signal and the estimate as random variables makes this quantity a random variable; the (random) bias is then

$$E_D[Q_q(X - \hat{X} + D) - (X - \hat{X} + D)].$$

To deal more easily with this expression, we introduce quantities which behave better than X and $\hat{X}$. Let $R = X - \hat{X} - Q_q(X - \hat{X})$. Note that R is the difference between a sample and its estimate, translated by a multiple of q to be in the range $[-q/2, q/2]$. If no dither is used, R is the error in the reconstructed sample, and when a dither signal is used, R should still be distributed in the same way as the no-dither reconstruction error (that is, ideally uniformly distributed over $[-q/2, q/2]$ but in practice typically slightly peaked at 0). Let $\tilde{R} = Q_q(R + D) - D$, so that $\tilde{R}$ is the quantized value of subtractively dithered R. Note that if no dither is used then $\tilde{R} = 0$.

With these definitions we have

$$\begin{aligned} E_D[\tilde{R} - R] &= E_D[Q_q(R + D) - D - (X - \hat{X}) + Q_q(X - \hat{X})] \\ &= E_D[Q_q(X - \hat{X} - Q_q(X - \hat{X}) + D) - (X - \hat{X} + D) + Q_q(X - \hat{X})] \\ &= E_D[Q_q(X - \hat{X} + D) - (X - \hat{X} + D)], \end{aligned}$$

so $E_D[\tilde{R} - R]$ expresses the sample bias in terms of R and D. We let $A(P_D)$ be the mean-squared value of this bias (averaged over R); that is

$$A(P_D) = E_R[(E_D[\tilde{R} - R])^2],$$

or, equivalently,

$$A(P_D) = \int_{[-\frac{q}{2}, \frac{q}{2}]} (r - E[\tilde{R}|R = r])^2 dP_R(r). \tag{1}$$

We primarily consider the case where R is distributed uniformly on $[-q/2, q/2]$ (or at least where we weight the bias uniformly over this range of R) so that

$$A(P_D) = \int_{-\frac{q}{2}}^{\frac{q}{2}} (r - E[\tilde{R}|R = r])^2 dr. \tag{2}$$

This metric is discussed in [4] and [6]. Intuitively, $A(P_D)$ indicates the degree to which the expected mean of a sample can be biased by the quantization reference point.

Suppose $A(P_D)$ is given by (2) and $C(P_D)$ is the variance or the second moment of P_D. When no dither is used, $C = 0$ and $A = q^2/12$. When a dither that is uniform on $[-q/2, q/2]$ is used, $C = q^2/12$ and $A = 0$. Note that 0 is the minimum value of both $A(P_D)$ and $C(P_D)$. Our definitions of A and C are intended for comparison among dither distributions when q is constant. There is no obvious significance of a specific value of A or C, so we can only determine the range of best compromises between A and C. In a particular application, experimentation and subjective judgement will be needed to determine which of these "best compromises" to use.

3 Overview of Results

In Section 4 we present several results concerning the nature of optimal dither distributions. In particular, we show (Theorem 5) that when $A(P_D)$ is given by (2) and $C(P_D)$ is either the variance or the second moment of P_D, then the optimal trade-off between $C(P_D)$ and $A(P_D)$ occurs for dither distributions which are uniform on $[-k/2, k/2]$ for $k \in [0, q]$ (where $k = 0$ corresponds to no dither).

The discrete case is also addressed. In that case the samples are integers, $q = 2\delta+1$ where δ is the maximum absolute error allowed, and D must take on integer values. When $A(P_D)$ is the discrete analogue of (2) and $C(P_D)$ is the second moment of P_D, then the optimal distributions are those which are uniform on $\{-k, \ldots, k\}$ where $k \in \{0, \ldots, \delta\}$, or are a convex combination of two such distributions with consecutive values of k (Theorem 9).

As an example of the use of these results, an image we refer to as "munar" was compressed with a simple predictive algorithm with maximum sample error $\delta = 2$. A dither signal uniformly distributed on $\{-k, \ldots, k\}$ was used, where $k = 0$, 1, or 2. Note that $k = 0$ corresponds to no dither and $k = 2$ corresponds to standard subtractive dither. Table 1 demonstrates how the cost of dithering increases as the dither signal amplitude increases. Figure 1 contains a portion of the original "munar" image, and Figure 2 shows an enlarged and contrast-enhanced detail area of the original and reconstructed images, from which it can be seen how the dither signal amplitude affects the appearance of the reconstructed image. We make no claim as to which version is "best". When displayed normally, the original and all three reconstructed images are virtually indistinguishable; however, near-lossless compression is appropriate for images which will be subject to scientific analysis, in which case the appearance of these enhanced images is quite relevant.

4 Detailed Results

In this section we formally state theorems indicating that the dither distributions described in the previous section are optimal as claimed. Along the way we present general results which could simplify the process of determining optimal distributions if the basic assumptions are modified. Proofs or sketches of proofs may be found in Appendix C of [2].

k	rate (bits/pixel)	RMSE	Error distribution (percentage of pixels)				
			-2	-1	0	1	2
0	2.819	1.370	18.6	21.2	22.1	20.1	18.0
1	2.842	1.405	19.8	20.2	20.5	20.1	19.5
2	2.873	1.414	19.9	20.0	19.9	20.1	20.0

Table 1: Result of compressing "munar" with maximum pixel error $\delta = 2$ and a dither distribution uniform on $\{-k, \ldots, k\}$. For comparison, the lossless compression rate obtained with the same algorithm was 5.049 bits/pixel. Notice that the error distribution becomes more uniform as the dither signal amplitude increases.

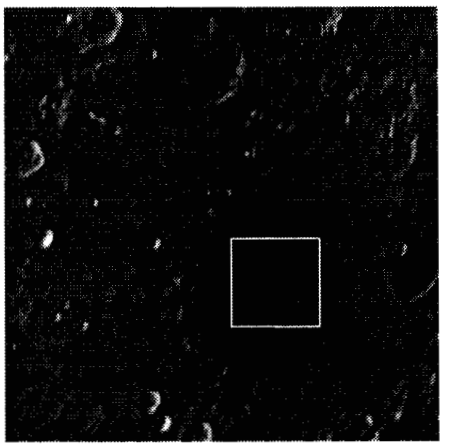

Figure 1: A portion of the "munar" image used in our dither example. The detail area shown in Figure 2 is indicated.

4.1 Continuous Case

We first consider the case where R and D may take on a continuum of values. We scale the numbers involved so that $q = 1$.

Although our choices of $A(P_D)$ and $C(P_D)$ under which we find optimal dither distributions are reasonable, it is possible that a more detailed analysis of a particular situation may yield more refined functions $A(P_D)$ and $C(P_D)$. We have not analyzed any specific alternate formulation in detail, but we present some results which may be helpful in determining optimal dither distributions for alternate formulations. Theorems 1, 3 and 4 give cases in which it is sufficient to consider distributions which are concentrated on certain intervals or which are symmetric. Theorem 2 gives conditions guaranteeing the existence of optimal dither distributions.

For these results, we assume $A(P_D)$ is of the form (1) but we no longer assume P_R is uniform on $[-\frac{1}{2}, \frac{1}{2}]$. We also do not assume that $C(P_D)$ is the variance of D or the second moment of D. However, we do place several restrictions on C. We assume that the cost of no dither is 0, and that $C(P_D) \geq 0$ for each P_D. The cost function

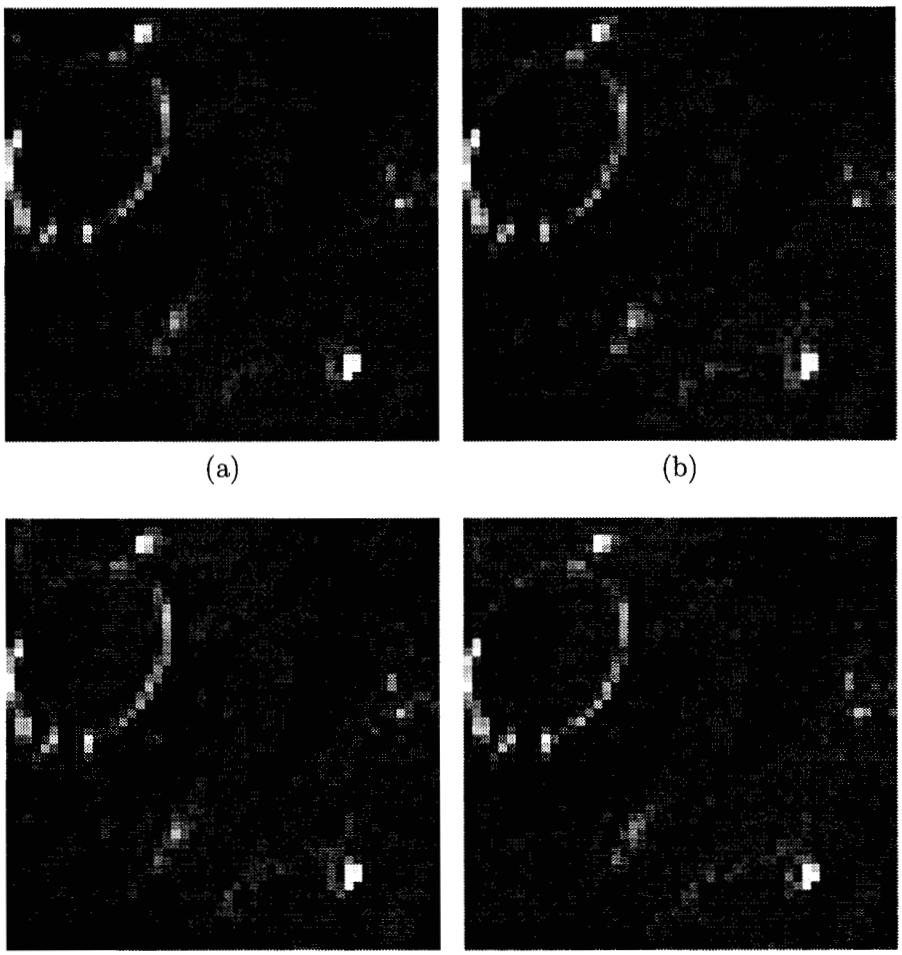

Figure 2: Dither example with maximum pixel error $\delta = 2$. Each image is magnified and contrast-enhanced. (a) Detail of original, (b) compressed and decompressed with $k = 0$ (no dither), (c) $k = 1$, and (d) $k = 2$ (standard dither). As k increases from 0 to 2, the appearance of streaks and artificial regions of constant intensity decreases, but an overall grainy look becomes more evident.

must also be one of the following two types:

Type A: For any (measurable) function $f : \mathbf{R} \to \mathbf{R}$ satisfying $0 \leq |f(x)| \leq |x|$ for all real x, and for any P_D, we require $C(P_{f(D)}) \leq C(P_D)$,

Type B: The cost function satisfies the following:

1. The cost is invariant to translations of D; that is, for any x we have $C(P_{D+x}) = C(P_D)$.
2. For any (measurable) function $f : \mathbf{R} \to \mathbf{R}$ satisfying $0 \leq |f(x)| \leq |x|$ for all real x, and for any P_D with mean 0, we require that $C(P_{f(D)}) \leq C(P_D)$.

Intuitively, a Type A cost function must assign higher cost to distributions which have larger amplitudes. A Type A cost function must be symmetric about 0 in the sense that $C(P_D) = C(P_{-D})$, since with $f(x) = -x$ we have $C(P_D) \leq C(P_{-D}) \leq C(P_D)$. The second moment of D is a Type A cost function.

A Type B cost function must be symmetric about μ when restricted to distributions with mean μ. The variance of D is a Type B cost function.

Theorem 1 *For any $P_{D'}$, there exists a $P_{D''}$ with $C(P_{D''}) \leq C(P_{D'})$ and $A(P_{D''}) = A(P_{D'})$, and*

(i) if C is a Type A cost function, then $P_{D''}$ can be chosen to be concentrated on $[-\frac{1}{2}, \frac{1}{2}]$;

(ii) if C is a Type B cost function, then $P_{D''}$ can be chosen to be concentrated on an interval of the form $[c - \frac{1}{2}, c + \frac{1}{2}]$, where $c \in [-\frac{1}{2}, \frac{1}{2}]$;

(iii) if C is a Type B cost function which is continuous under the topology of weak convergence[1], then $P_{D''}$ can be chosen so that its mean μ is in $[-\frac{1}{2}, \frac{1}{2}]$ and $P_{D''}$ is concentrated on $[\mu - \frac{1}{2}, \mu + \frac{1}{2}]$.

Note that Theorem 1 implies that in all cases which obey our general restrictions on the cost function, it suffices to consider dither distributions which are concentrated on $[-1, 1]$.

Theorem 2 *If the (Type A or B) cost function $C(P_D)$ is continuous under the topology of weak convergence and P_R can be described by a density function p_R (that is, $P_R(S) = 0$ whenever the Lebesgue measure of S is 0), then for each $\alpha \geq 0$ there exists a P_D^* which minimizes $A(P_D)$ subject to $C(P_D) \leq \alpha$.*

Theorem 3 *Suppose P_R is symmetric about 0 and can be described by a density function p_R. Suppose also that C is a Type A cost function and C is convex $\cup$. Then for any $P_{D'}$, there exists a $P_{D''}$ which is symmetric about 0, is concentrated on $[-\frac{1}{2}, \frac{1}{2}]$, and for which $A(P_{D''}) \leq A(P_{D'})$ and $C(P_{D''}) \leq C(P_{D'})$.*

[1] A good reference for the topology of weak convergence is Appendix III of [1]

Observe that Theorem 3 implies that when its hypothesis holds, it suffices to consider dither distributions which are symmetric about 0 (and thus have mean 0) and are concentrated on $[-\frac{1}{2}, \frac{1}{2}]$.

Theorem 4 *Suppose P_R is the uniform distribution on $[-\frac{1}{2}, \frac{1}{2}]$. Suppose also that $C(P_D)$ is a Type B cost function and that $C(P_D)$ is convex $\cup$ when restricted to P_D with mean 0. Then for any $P_{D'}$, there exists a $P_{D''}$ which is symmetric about 0, is concentrated on $[-\frac{1}{2}, \frac{1}{2}]$, and for which $A(P_{D''}) \leq A(P_{D'})$ and $C(P_{D''}) \leq C(P_{D'})$.*

Finally, we have the basic continuous case:

Theorem 5 *Suppose $A(P_D)$ is given by (2) and $C(P_D)$ is either the variance or the second moment of P_D. Suppose $k \in [0, q]$ and let P_D^* be the uniform distribution on $[-k/2, k/2]$. Then P_D^* minimizes $A(P_D)$ subject to $C(P_D) = k^2/12$.*

This result is proved by observing that $A(P_D)$ is convex $\cup$ and then showing that for any zero mean P_D with the same cost as P_D^*,

$$\frac{\partial}{\partial \epsilon} A((1-\epsilon)P_D^* + \epsilon P_D)\Big|_{\epsilon=0} \geq 0.$$

4.2 Discrete Case

For the discrete case we assume that R and D take on integer values. We consider only the case where q is odd, with $q = 2\delta + 1$. We use P_D to denote the (discrete) dither probability distribution, with $P_D(i) = \Pr(D = i)$. Much of the notation is the similar to that of the continuous case and we rely on context to distinguish the two. In the basic discrete case $A(P_D)$ is defined as

$$A(P_D) = \frac{1}{2\delta + 1} \sum_{r=-\delta}^{\delta} (r - E[Q_q(r + D) - D])^2 \tag{3}$$

and the cost $C(P_D)$ is the second moment of D.

Again we remove some of our specific assumptions on $A(P_D)$ and $C(P_D)$. We consider $A(P_D)$ of the form

$$A(P_D) = \frac{1}{2\delta + 1} \sum_{r=-\delta}^{\delta} (r - E[Q_q(r + D) - D])^2 P_R(r),$$

where P_R is the (discrete) residual distribution. We require the cost of no dither to be 0 and that $C(P_D) \geq 0$ for each P_D. The cost function must also be one of the following two types:

Type A For any function $f : \mathbf{Z} \to \mathbf{Z}$ satisfying $0 \leq |f(i)| \leq |i|$ for all integers i, and for any P_D, we require $C(P_{f(D)}) \leq C(P_D)$.

Type B The cost function satisfies the following:

1. The cost is invariant to translations of D; that is, for any integer i we have $C(P_{D+i}) = C(P_D)$.
2. For any function $f : \mathbf{R} \to \mathbf{R}$ satisfying $0 \leq |f(x)| \leq |x|$ and $f(x) - x \in \mathbf{Z}$ for all real x, and for any P_D, we require that $C(P_{f(D-\mu)+\mu}) \leq C(P_D)$, where μ is the mean of D.

Intuitively, a Type A cost function must assign higher cost to distributions which have larger amplitudes. As is the continuous case, a Type A cost function must be symmetric about 0 in the sense that $C(P_D) = C(P_{-D})$, since with $f(i) = -i$ we have $C(P_D) \leq C(P_{-D}) \leq C(P_D)$. The second moment of D is a Type A cost function.

A Type B cost function must be symmetric about 0 when restricted to distributions with mean 0. The variance of D is a Type B cost function.

Theorem 6 *For any $P_{D'}$, there exists a $P_{D''}$ with $C(P_{D''}) \leq C(P_{D'})$ and $A(P_{D''}) \leq A(P_{D'})$, and*

(i) if C is a Type A cost function, then $P_{D''}$ can be chosen to be concentrated on $\{-\delta, \ldots, \delta\}$;

(ii) if C is a Type B cost function, then $P_{D''}$ can be chosen to be concentrated on a set of the form $\{c - \delta, \ldots, c + \delta\}$, where $c \in \{-\delta, \ldots, \delta\}$.

Note that Theorem 6 implies that in all cases which obey our general restriction on the cost function, it suffices to consider dither distributions which are concentrated on $\{-(q-1), \ldots, q-1\}$.

Theorem 7 *If $C(P_D)$ is continuous when restricted to P_D which are concentrated on $\{-(q-1), \ldots, q-1\}$, then for each $\alpha \geq 0$ there exists a P_D which minimizes $A(P_D)$ subject to $C(P_D) \leq \alpha$.*

Theorem 7 is the discrete analogue of Theorem 2.

Theorem 8 *Suppose P_R is symmetric about 0 and C is a Type A cost function which is convex $\cup$. Then for any $P_{D'}$, there exists a $P_{D''}$ which is symmetric about 0, is concentrated on $\{-\delta, \ldots, \delta\}$, and for which $A(P_{D''}) \leq A(P_{D'})$ and $C(P_{D''}) \leq C(P_{D'})$.*

Theorem 8 is the discrete analogue of Theorem 3.

Note that there is no simple discrete analogue of Theorem 4. Our results for the basic discrete case (below) apply when the cost function is the second moment of P_D and do not apply when the cost function is the variance of P_D.

Theorem 9 *Suppose $A(P_D)$ is given by (3) and $C(P_D)$ is the second moment of P_D. Suppose P_D^* is the discrete distribution given by*

$$P_D^*(i) = \begin{cases} \alpha & \text{if } |i| < k \\ \frac{1+\alpha}{2} - k\alpha & \text{if } |i| = k \\ 0 & \text{if } |i| > k, \end{cases}$$

where $k \in \{1, \ldots, \delta\}$ and $\alpha \geq \frac{1+\alpha}{2} - k\alpha$. Then P_D^* minimizes $A(P_D)$ subject to $C(P_D) = C(P_D^*)$.

Note that this P_D^* is a convex combination of the uniform distributions on $\{-k, \ldots, k\}$ and $\{-(k-1), \ldots, k-1\}$.

Theorem 9 is proved by first noting that $A(P_D)$ is a convex $\cup$ function and then using the Kuhn-Tucker conditions [3] to show optimality of P_D^*.

5 Conclusion

We have shown that it is possible to compromise between the degree that artifacts are present in a near-losslessly compressed signal and the rate and distortion costs of subtractive dithering. Under reasonable assumptions, the optimal dither distributions are shown to be uniform distributions on the interval $[-k/2, k/2]$, where $k \in [0, q]$. Less specific results were also presented under alternate assumptions.

A possible direction for future work on this subject is to consider how dither should be used when the quantizer is not uniform. Of particular interest is the case where the center quantization "bin" is larger than the others, since such a quantizer can yield an improvement is compression under some circumstances.

References

[1] Partick Billingsley. *Convergence of Probability Measures.* John Wiley & Sons, Inc., New York, 1968.

[2] M. Klimesh. Quantization considerations for distortion-controlled compression. *The Telecommunications and Mission Operations Progress Report 42-139, July–September 1999*, Jet Propulsion Laboratory, Pasadena, California. November 15, 1999. http://tmo.jpl.nasa.gov/tmo/progress_report/42-139/139K.pdf

[3] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proc. 2nd Berkeley Symposium on Mathametical Statistics and Probability*, pages 481–492, Berkeley, 1951. University of California Press.

[4] Lawrence Gilman Roberts. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8:145–154, February 1962.

[5] Daniël W. E. Schobben, Rob A. Beuker, and Werner Oomen. Dither and data compression. *IEEE Transactions on Signal Processing*, 45(8):2097–2101, August 1997.

[6] Leonard Schuchman. Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology*, COM-12:162–165, December 1964.